Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances

Pierre Zweigenbaum

Service d'informatique médicale, Assistance publique - Hôpitaux de Paris
& Département de biomathématiques, Université Paris 6

1 Introduction

Les systèmes d'information médicaux doivent se distinguer par leur capacité à enregistrer et transmettre des informations et des connaissances médicales. Les objectifs de ces informations sont variés ; citons en particulier [1] le soin au patient, l'évaluation de la qualité, la recherche et l'épidémiologie, la planification et la gestion, sans oublier la formation.

Le moyen naturel et habituel pour exprimer ces informations est la *langue naturelle*, encore appelée *texte libre*. Elle constitue le moyen le plus simple et le plus universel pour décrire des faits, exprimer des observations, transmettre des connaissances, avec le degré de précision ou d'imprécision souhaité. Les dossiers de patients aussi bien que les cours de médecine ou les références médicales opposables sont ainsi rédigés en langue naturelle.

La puissance de la langue naturelle crée en même temps un obstacle à son utilisation pour le traitement de l'information. De fait, les systèmes informatiques éprouvent des difficultés en présence de la paraphrase ou de la construction de nouveaux concepts, omniprésents dans l'emploi de la langue. Ils ont tendance à buter sur l'ambiguïté de certains énoncés, pourtant clairs dans le contexte dans lequel ils ont été écrits. En résumé, ils manipulent des symboles formels, et ne peuvent de ce fait appréhender directement des textes (voir cependant l'article de Pierre Dujols dans ce numéro). Pour traiter l'information médicale avec une machine, il faut lui en fournir un modèle formel.

L'idée générale est de donner à un système informatique les moyens d'effectuer des calculs sur une *représentation* de l'information médicale : au minimum, il devra pouvoir comparer deux représentations et déterminer si elles sont totalement identiques, partiellement identiques ou tout à fait différentes. La base d'une telle représentation est un inventaire normalisé des éléments d'information que l'on veut manipu-

ler : l'ensemble des *concepts* pertinents pour le domaine abordé, et celui des *relations* qui permettent de relier des concepts généraux à des concepts plus spécifiques, ou de construire des concepts complexes à partir de concepts plus simples.

Après avoir passé en revue différents problèmes qui se posent dans l'échange d'information (section 2), nous examinons ici les principales approches existantes de modélisation des concepts médicaux. Ces approches vont de la constitution de référentiels de codage précoordonnés (par exemple, la Classification internationale des maladies ; section 3) à la mise au point de systèmes formels génératifs permettant de composer à l'infini les concepts élémentaires d'une ontologie (par exemple, celle du projet GALEN ; section 4), plus puissants mais en même temps plus complexes à utiliser. Nous concluons sur la tension induite entre le besoin d'expressivité et la recherche de la normalisation et de la formalisation (section 5).

2 De la difficulté d'échanger des informations

Pour que la communication entre deux personnes ou systèmes informatiques fonctionne, il faut que le contenu du message produit par l'*émetteur* soit préservé lorsque le message est interprété par le *récepteur* (voir par exemple [2]). Une condition souhaitable est bien sûr que le message soit transmis au récepteur dans son intégrité. Pour les systèmes informatiques, cette condition est assurée par les six premières couches (sur sept au total) du modèle ISO-OSI (voir par exemple [3]) qui régit les échanges d'information électroniques. La mise en application de ce modèle assure entre autres les fonctions de routage du message dans un réseau, ainsi que son encodage initial et son décodage à l'arrivée.

Une fois le message arrivé et restitué dans sa forme initiale, encore faut-il que le récepteur, qu'il soit un être humain ou un système informatique, lui donne le « même sens » que l'émetteur. La question générale en jeu ici, pour la communication entre systèmes informatiques, est celle de l'*interopérabilité sémantique* [4]. Une application informatique qui reçoit ce message et qui va effectuer des calculs (par exemple, statistiques) à partir de l'information qu'il véhicule doit pouvoir mettre en rapport son contenu avec ses propres représentations.

Parmi les facteurs qui peuvent rendre difficile la lecture d'un message, l'*ambiguïté* est le plus souvent cité. Elle correspond, de façon générale, à une situation dans laquelle un message peut être interprété de plusieurs façons. Ce terme recouvre plusieurs réalités.

– Le *manque de consensus* sur la définition d'une notion est source d'ambiguïté. Cette notion peut recouvrir des réalités différentes d'un lieu à un autre, ou peut varier dans le temps du fait de l'évolution des connaissances ou des techniques médicales. Sans qualification supplémentaire, elle ne pourra pas être interprétée de façon certaine.

– L'emploi d'un *mot possédant plusieurs sens* (*polysème*) peut créer une ambiguïté

accidentelle. Ainsi, *« genou »* ne désigne pas le même objet dans *« une occlusion de la coronaire droite au genou inférieur »* et *« une ostéotomie du genou droit pour arthrose »*.

– L'*imprécision* correspond à une description insuffisamment spécifiée pour permettre d'identifier une notion utile dans un contexte donné. Par exemple, *« infarctus »*, sans mention de localisation anatomique et hors d'un contexte spécifique, sera en général considéré comme imprécis.

Une autre difficulté potentielle est la possibilité de désigner une même notion de plusieurs façons. Deux termes peuvent être considérés comme *synonymes* : par exemple, *« spondylarthrite ankylosante »* et *« pelvispondylite rhumatismale »*. En présence d'expressions plus larges, on parlera plus généralement de *paraphrase*. La synonymie est une autre manifestation de la variabilité de dénomination des notions médicales dans l'espace et dans le temps. Elle est en général relative, dans la mesure où des dénominations différentes se distinguent souvent par une nuance que l'on peut négliger dans un contexte fixé, mais qui pourrait avoir son importance dans un contexte particulier. Une situation commune où la synonymie se manifeste est celle où l'on emploie deux systèmes qui se fondent sur des terminologies différentes. Il est alors fréquent qu'une même notion soit identifiée par deux termes différents. Ainsi, le terme *« prothèse oculaire, SAI »* (nomenclature SNOMED) est désigné dans le thésaurus MeSH par *« œil artificiel »* (ces deux terminologies sont discutées plus loin).

Un fait, une observation que l'on veut enregistrer ou transmettre peut être plus ou moins complexe, et être décomposable en plusieurs notions plus simples. Par exemple, si aucune dénomination conventionnelle n'est répertoriée pour le fait *« hémorragie digestive »*, mais que l'on en a une pour *« hémorragie »* et pour *« voies digestives »*, on pourra chercher à exprimer ce fait en *composant* ces deux expressions plus simples (par exemple, sur le mode *« hémorragie* des voies digestives »). Cette possibilité de composition est un mécanisme naturel et fondamental de la langue, mais qui demande un effort supplémentaire d'interprétation lors de la réception du message. D'une part, il n'est pas toujours facile ni même possible pour un système informatique de retrouver le « sens » d'une expression composite formée à partir d'expressions connues plus simples (voir par exemple [5]). D'autre part, même pour une personne, cette opération requiert en général une connaissance minimale du domaine dont on parle : dans *« hémorragie des voies digestives »*, *« des »* signifie *« localisée dans les »* ; et dans *« hémorragie digestive »*, *« digestive »* signifie *« localisée dans les voies digestives »*. Cette *(non-)compositionnalité* est une difficulté importante de la représentation d'informations et de connaissances un tant soit peu complexes. La façon la plus simple de s'en affranchir est de prévoir à l'avance une dénomination spécifique pour chacune des notions qui pourront être représentées. Mais on limite alors a priori le nombre de ces notions à une quantité relativement faible (couramment de l'ordre de la dizaine de milliers, comme dans la Classification internationale des maladies), ce qui contraint fortement ce que l'on peut exprimer.

On le voit, la représentation que l'on choisit d'adopter pour enregistrer les informations et connaissances médicales conditionne la façon dont ces problèmes vont être abordés. Son fondement est une *modélisation des concepts médicaux* pertinents pour l'objectif de communication fixé. Nous examinons dans un premier temps les approches terminologiques, qui sous-tendent la plupart des systèmes de codage effectivement utilisés (section 3). Nous étudions dans un second temps les méthodes compositionnelles et formelles, d'une plus grande expressivité mais en même temps plus complexes, qui font l'objet des principaux développements actuels (section 4).

3 Approches terminologiques et normalisation de l'expression des concepts médicaux

Les formalisations du sens se réfèrent à une vision du monde caractérisée par les trois sommets d'un *triangle sémiotique* (voir entre autres [6, 7, 8]).

- On suppose que l'on peut identifier des *objets*, concrets ou abstraits, dans le monde (par exemple, *« cœur »*, *« artériographie »*, *« raison »*). C'est à propos de ces objets que l'on veut transmettre des informations ou exprimer des connaissances.
- On appréhende les objets en s'en faisant une idée, en les idéalisant sous forme de *concepts*.
- On parle des concepts ou des objets à l'aide d'énoncés dans une langue. On suppose qu'un concept ou un objet pourra être exprimé par une expression, un *terme* de cette langue.

Cette triade reste certes un objet de débat ([9], voir aussi la section 5.2). Elle n'en forme pas moins le fondement de la doctrine terminologique créée dans les années trente par Wüster au sein du cercle de Vienne, et l'on peut considérer que les formalisations de l'information médicale héritent de cette tradition. Nous précisons d'abord la notion de *terminologie*, puis discutons celles de thésaurus, classification et nomenclature.

3.1 Terminologie

La terminologie, en tant que science, s'intéresse au recensement des concepts d'un domaine et des termes qui les désignent. Dans le reste de cet article, nous désignons par *« terminologie »* le produit de cette activité. Les terminologies s'intéressent essentiellement (et historiquement) à des domaines techniques, et visent à faciliter l'échange de connaissances dans une langue et d'une langue à l'autre. Pour cela, on va normaliser l'expression des concepts du domaine en fixant les *termes* qui les désignent. On va de plus rendre compte dans la terminologie de l'agencement relatif des concepts recensés. Ces concepts peuvent être reliés par des *relations*, en particulier de spécialisation

- généralisation. Ainsi, *« infarctus aigu du myocarde »* est un concept plus spécifique que *« infarctus du myocarde »*, lui-même plus spécifique que *« maladie cardiaque »*, dans la mesure où un infarctus du myocarde est par essence une maladie cardiaque. Une *définition* de chaque concept est souvent fournie. Une définition typique distingue un concept du concept plus général le plus proche (son *genre proche*) en énonçant ses *différences spécifiques*. Une terminologie modélise ainsi un système de concepts sous la forme d'un système de termes normalisés.

En donnant la primauté au concept, donnée du domaine, sur le terme, désignateur conventionnel du concept, l'aspect multilingue est considérablement simplifié. Chaque concept peut être désigné par un terme propre à chaque langue. Une terminologie multilingue postule ainsi que les concepts d'un domaine sont communs à toutes les langues considérées. Pour un domaine donné, on aura donc un système de concepts unique, reflété dans chaque langue par un jeu de termes approprié. Les terminologies multilingues constituent une aide précieuse pour les traducteurs de textes techniques et scientifiques.

La plupart des terminologies ont une visée normative. De fait, l'emploi des termes normalisés d'une terminologie de référence résoud la plupart des difficultés d'échange d'information mentionnées ci-dessus (section 2).

L'ambiguïté est énormément réduite, si ce n'est supprimée. Par définition, une terminologie de référence spécifie une norme pour le domaine considéré. On sait ainsi dans quel sens chaque terme est employé. De plus, lors de la constitution d'une terminologie, les mots polysémiques seront soit évités, soit précisés par d'autres mots, en veillant à ce que chaque terme désigne un concept unique du domaine. Par exemple, le sens de *« sinus »* sera différencié dans *« sinus paranasal »* et *« sinus pilonidal »*. L'imprécision, si elle n'est pas supprimée, peut être quant à elle mieux encadrée. D'une part, dans le contexte d'un domaine suffisamment spécifique, cette imprécision est moins fréquente. D'autre part, la possibilité de hiérarchiser les concepts permet de relier explicitement un terme générique, et de ce fait imprécis (p.ex., *« infarctus »*), aux termes plus spécifiques qui peuvent le préciser (p.ex., *« infarctus du myocarde »*, *« infarctus pulmonaire »*).

La synonymie ou paraphrase correspond à une situation dans laquelle un concept unique serait désigné par plusieurs termes différents. L'adhésion à une terminologie de référence supprime ce cas de figure, chaque concept se voyant associer un terme unique, normalisé. Pour faciliter l'établissement de correspondances entre des termes courants et les termes normalisés, certaines terminologies incluent des termes supplémentaires pour désigner un même concept : des synonymes du terme normalisé. On conserve ainsi une possibilité de synonymie, mais elle est encadrée a priori.

Les terminologies ont en général une approche de recensement, de compilation des concepts d'un domaine et de leurs termes associés. La composition de plusieurs concepts pour en former d'autres plus complexes n'est habituellement pas considérée comme faisant partie de son champ. En revanche, la description de relations (spécialisation - généralisation, tout - partie, etc.) entre concepts, en particulier entre un concept complexe et d'autres concepts plus élémentaires, est un pas vers une décomposition

partielle et figée des concepts. Nous revenons sur ce point à la section 4.

Notons pour finir que l'on peut également identifier un concept par un *code*, par exemple numérique ou alphanumérique. Dans le cadre d'une terminologie où aucun terme ne peut désigner plusieurs concepts à la fois, ce code est théoriquement redondant avec le terme préférentiel. Ces identifiants alphanumériques, historiquement plus pratiques comme données informatiques, sont souvent utilisés pour encoder la hiérarchie des concepts, essentiellement en jouant sur le nombre plus ou moins grand de caractères dans le code (l'ajout de caractères supplémentaires correspond à un concept plus spécifique ou constituant une partie du concept précédent).

3.2 Des terminologies différentes pour des objectifs distincts

Différents objectifs de traitement de l'information médicale ont amené à constituer des terminologies de natures différentes [10]. Nous examinons les cas de la recherche d'information (thésaurus), du recueil de données à des fins statistiques (classification) et de la description d'observations cliniques (nomenclature). Notons qu'il est difficile de donner des définitions consensuelles de ces notions, ce qui est un comble dans ce domaine : il semble que même différents sous-comités de l'ISO en proposent des définitions différentes [11, p. 554]. Nous reflétons donc plutôt ici l'usage habituel qui en est fait dans le domaine médical. On pourra se reporter à divers ouvrages et articles pour des éclairages croisés sur ces notions [2, 1, 10, 11].

3.2.1 Recherche d'information, thésaurus

La *recherche d'information*, ou *recherche documentaire*, a pour but d'identifier les documents contenant des informations répondant à une requête initiale. Les deux applications les plus connues de cette technique sont la recherche bibliographique dans des bases d'articles scientifiques (par exemple, Medline ou Pascal) et la recherche en texte intégral sur Internet. Nous nous intéresserons à la première, qui fait appel à une terminologie contrôlée pour indexer les documents : un *thésaurus*. La base Medline, la plus employée dans le domaine biomédical, utilise le thésaurus MeSH (*Medical Subject Headings*, [12]). Le principe de l'indexation, effectuée manuellement par des indexeurs professionnels, est de décrire un article par les thèmes principaux dont il traite, ces thèmes étant choisis parmi ceux recensés dans le thésaurus. La recherche d'un article, pour un utilisateur, se fait ensuite en mentionnant ses thèmes d'intérêt, et les documents indexés par ces concepts seront retrouvés.

Les concepts inclus dans un thésaurus de recherche d'information sont choisis pour couvrir le domaine avec un degré de finesse qui dépend de l'effort consenti pour l'indexation. On peut sans doute considérer le MeSH comme un exemple de thésaurus à grain fin. Les termes (« *descripteurs* », ou encore « *vedettes* ») employés pour désigner les thèmes du domaine ne sont pas nécessairement des expressions employées effectivement dans les documents. On trouve par exemple dans le MeSH le descripteur « *vaisseaux coronaires, maladies* » plutôt que l'expression plus naturelle « *maladies*

des vaisseaux coronaires », ou encore *« infarctus myocarde »* plutôt que *« infarctus du myocarde »*. Pour faciliter la formulation de requêtes, les descripteurs sont souvent accompagnés de synonymes. Enfin, les thésaurus incluent généralement des relations entre concepts : spécialisation - généralisation, tout - partie sont les plus fréquentes, ainsi que la relation générale de voisinage *« lié à »*, qui permet de rechercher des documents traitant de notions proches.

3.2.2 Recueil orienté de données, classification

La description d'informations peut être liée à un objectif précis d'observation, correspondant à une question spécifique qui guide le recueil de données [10]. C'est par exemple le cas du recueil de diagnostics à des fins de santé publique ou d'évaluation de l'activité hospitalière. Le système de concepts que l'on va mettre en place pour représenter les réponses possibles à cette question est directement influencé par cet objectif. Pour pouvoir effectuer des calculs statistiques sur les données recueillies, on va partitionner l'espace des réponses en classes, de préférence statistiquement équilibrées. Ces classes constituent une *classification*. Elles doivent couvrir l'ensemble des réponses possibles. Leur granularité dépend des objectifs poursuivis. La définition de classes plus spécifiques, partitionnant elles-mêmes les classes plus générales, hiérarchise la classification. Elle permet de travailler à différents niveaux de granularité. La *Classification statistique internationale des maladies et des problèmes de santé connexes* (CIM, [13]) est un exemple de classification hiérarchique.

En reprenant notre modèle terminologique, les concepts d'une classification sont ses classes. Les termes d'une classification (aussi appelés *rubriques*) sont souvent des expressions d'un *métalangage* plutôt que les expressions que l'on pourrait trouver dans des textes naturels. Ils constituent alors des instructions guidant le choix d'une classe à laquelle affecter un cas donné. Les expressions telles que *« sans autre indication »* (SAI), *« Autres ... »* ou *« à l'exclusion de »* sont typiques de ce métalangage.

3.2.3 Recueil ouvert de données, nomenclature

Lorsque le but est de décrire des informations cliniques le plus précisément et fidèlement possible, les classifications telles que définies ci-dessus, trop orientées vers un objectif précis, se révèlent peu adaptées. On a en effet besoin de disposer d'une terminologie fournissant un éventail plus varié et plus précis de concepts médicaux. La notion de *nomenclature* est une autre variante de la notion générale de terminologie introduite à la section 3.1. Elle vise à recenser tous les concepts d'un domaine, sans se restreindre a priori à un objectif spécifique.

La *Nomenclature systématique des médecines humaine et vétérinaire* (SNOMED [14]) est une *nomenclature systématique multiaxiale* : elle permet de projeter les concepts médicaux selon plusieurs *axes orthogonaux*. Les huit axes principaux de cette projection sont Topographie (T), Morphologie (M), Fonction (F), Organismes vivants (L), Médicaments, produits chimiques et biologiques (C), Agents, activités physiques

et forces naturelles (A), Métiers et professions (J), et Contexte social (S). La variété de ces axes distingue cette nomenclature d'une classification monoaxiale comme la CIM, qui se limite essentiellement à un type de concept : les diagnostics. Par ailleurs, chaque axe est lui-même hiérarchisé, les concepts de différents niveaux étant liés par des relations de spécialisation (p.ex., « *brûlure* » (M-11100) est une sorte de « *blessure thermique* » (M-11000) qui est une sorte de « *blessure* » (M-10000)) ou du tout à la partie (p.ex., la « *crosse de l'aorte* » (T-42300) est une partie de l'« *aorte* » (T-42000)). Comme dans le thésaurus MeSH, chaque concept est désigné par un terme préférentiel et éventuellement par des synonymes. Ici cependant, le terme préférentiel comme les synonymes sont dans la quasi-totalité des cas des expressions naturelles que l'on peut trouver dans un texte (quelques méta-termes, comme « *SAI* » ou « *Autres ...* », sont toutefois employés). Notons que la SNOMED inclut de plus l'axe particulier G, dont nous reparlons ci-dessous, et deux classifications, une pour les actes (P) et une pour les diagnostics (D : elle référence la CIM-9). On compte donc au total onze « axes » dans la version actuelle de la SNOMED (version III, aussi appelée *SNOMED Internationale*).

La répartition des concepts en plusieurs axes a pour but additionnel de permettre de composer un concept complexe en combinant des concepts élémentaires pris dans ces axes. L'axe des Qualificatifs et termes relationnels (G) contient des concepts supplémentaires servant à qualifier ces concepts ou à préciser leurs liens dans le concept complexe. Par exemple, une « *appendicite aiguë* » (exemple adapté de [15]) pourra être représentée par la combinaison des concepts « *inflammation, SAI* » (M-41000), « *aigu* » (G-A231), « *dans* » (G-C006), « *appendice vermiculaire, SAI* » (T-59200). Nous développons cette possibilité ci-dessous dans la section 4.

4 Approches compositionnelles et systèmes formels de représentation des connaissances

4.1 Approches compositionnelles

La possibilité de composer plusieurs concepts simples pour représenter un concept plus complexe est absente de la notion de terminologie telle que nous l'avons présentée plus haut (section 3.1). Pour chaque objet pertinent du domaine considéré, un concept doit être identifié dans la terminologie, et un ou plusieurs termes lui seront associés. Ce type d'approche est qualifiée de *précoordonnée* : chaque combinaison pertinente de concepts élémentaires doit être explicitement prévue à l'avance lors de la constitution de la terminologie. La Classification internationale des maladies est un exemple de terminologie précoordonnée.

Des possibilités de composition plus ou moins puissantes ont été associées à certaines terminologies, dites *post-coordonnées*. Nous les examinons dans le reste cette section en nous inspirant de la gradation proposée par Spackman et Campbell [15].

4.1.1 Concepts non différenciés

La *cooccurrence* ou conjonction de deux ou plusieurs concepts constitue une forme élémentaire de combinaison de ces concepts. On pourra ainsi décrire (ou rechercher) un patient ayant à la fois un « *myélome multiple* » et une « *hypercalcémie* ». La conjonction de diagnostics de la CIM (ou d'actes) dans le recueil de données du Programme de médicalisation du système d'information (PMSI) fait intervenir ce type de combinaison.

On peut reformuler cette cooccurrence comme l'emploi de l'opérateur booléen *et*. On augmente les possibilités de recherche en employant également les opérateurs *ou* et *non*. La recherche documentaire permet souvent d'employer ces opérateurs. La recherche dans la base Medline, fondée sur l'emploi du thésaurus MeSH, en est un exemple.

4.1.2 Axes orthogonaux

Lorsque les concepts médicaux sont répartis selon plusieurs axes orthogonaux, comme dans la nomenclature SNOMED (voir la section 3.2.3), leur conjonction peut prendre un sens différent. En effet, plutôt que d'énumérer, par exemple, différents diagnostics, on va pouvoir composer un diagnostic en spécifiant ses différentes facettes. Les facettes principales proposées par la version 2 de la SNOMED étaient les suivantes [15] : procédure (acte), topographie, morphologie, étiologie (cet axe a été remplacé dans la version actuelle, la SNOMED III), fonction, et maladie, ainsi qu'un champ *qualificatif informationnel*. Un diagnostic d'*« appendicite aiguë »* peut ainsi être représenté par la décomposition suivante :

QI	P	T	M	E	F	D
aigu G-A231		appendice vermiculaire, SAI T-59200	inflammation, SAI M-41000			

On voit la puissance de ce mécanisme de composition. Une approche précoordonnée devrait énumérer tous les diagnostics précis, par exemple toutes les inflammations possibles des différentes localisations anatomiques. Une approche compositionnelle, ou *post-coordonnée*, fournit les éléments utiles pour composer ces diagnostics à la demande. Si l'on combine les 5 880 termes de morphologie de la SNOMED avec ses 12 936 topographies (soit simplement les colonnes 3 et 4 du tableau), on a un espace potentiel de 76 millions de concepts — dont tous bien sûr ne sont pas médicalement sensés. Il faut de plus corriger ce chiffre en tenant compte non pas du nombre de termes (qui inclut ici les synonymes), mais de concepts. L'ordre de grandeur reste cependant le même, et peut être mis en regard des quelque 10 000 diagnostics de la CIM-10.

Comme mentionné plus haut, deux des « axes » de la SNOMED III, Diagnostics et Procédures, sont particuliers. Ils contiennent des concepts complexes qui peuvent se projeter sur les autres axes élémentaires de la nomenclature. Cette projection est

fournie pour une bonne partie de ces concepts. Ainsi, on trouve, en regard de *« appendicite, SAI »* (D5-46100), une référence à *« appendice vermiculaire, SAI »* (T-59200) et à *« inflammation, SAI »* (M-40000).

4.1.3 Relations explicites

Une correspondance directe entre axe et facette limite pourtant la puissance expressive de la composition. La relation qui existe entre le concept complexe décrit et un concept élémentaire qui le compose, pris par exemple dans l'axe F, n'est pas explicitée. Plusieurs relations différentes sont pourtant envisageables : l'effet sur le corps, mais aussi les circonstances, etc. Si l'on veut représenter un *« choc hypovolémique »* (effet sur le corps) lors d'un *« homicide »* (circonstances), comme ces deux concepts sont tous deux dans l'axe F, on ne peut pas les exprimer conjointement et distinguer leurs rôles dans l'approche du paragraphe précédent.

Pour y parvenir, il faut employer des *relations* explicites, comme *« cause »*, *« circonstances »*, *« effet sur le corps »*, *« site anatomique »*, *« anomalie »*, etc. On peut alors représenter un concept complexe par un petit réseau de concepts et de relations. Par exemple, notre *« appendicite aiguë »* sera décomposée selon le schéma suivant (les concepts sont encadrés par des boîtes rectangulaires et les relations par des parenthèses) :

[inflammation, SAI] → (localisé dans) → [appendice vermiculaire, SAI]
↘ (cours) → [aigu]

Les termes cliniques du système Read V3.1 mis au point en Grande Bretagne (appelé maintenant *« NHS Clinical Terms »*) sont structurés par un ensemble de relations (*attributs*) [16], et entrent donc dans ce cadre de représentation [15]. L'axe G de la SNOMED III contient des concepts qui pourraient jouer le rôle de relations : par exemple, *« dû à »* (G-C001), *« associé à »* (G-C002), *« dans »* (G-C006), *« traité par »* (G-C014), *« irradiant à »* (G-C040), etc. Cependant, l'ensemble de ces concepts « relationnels » n'est pas identifié en tant que tel : la nomenclature SNOMED III ne propose pas un jeu de relations explicites, ni un mode d'emploi de ces relations pour composer des concepts complexes.

4.1.4 Contraintes sur les relations

La donnée d'un jeu de relations correspond à une prise de position plus affirmée sur la façon de modéliser des concepts complexes à partir de concepts élémentaires. Cependant, sans précision supplémentaire, rien n'empêche d'employer ces relations dans un sens qui ne correspond pas à l'intention initiale du modélisateur. Par exemple, la relation *« localisé dans »* peut être employée pour relier une *« fonction »* à une *« localisation anatomique »* (comme dans *« appendicite »*) ; peut-elle aussi relier un

« acte » à une *« localisation anatomique »* (pour exprimer, par exemple, le sens d'*« appendicectomie »*) ? Il s'agit donc de normaliser aussi l'emploi des relations.

Deux directions peuvent être suivies. La première consiste à s'efforcer de normaliser le sens des relations de la même façon qu'on a pu le faire pour le sens des concepts. Une définition explicite et une hiérarchisation pourront participer à cet objectif. La seconde consiste à synchroniser relations et concepts, en énonçant des contraintes de combinaison des unes avec les autres. Par exemple, on pourra imposer que la relation *« localisé dans »* relie systématiquement un concept de type *« fonction »* ou *« morphologie »* à un concept de type *« localisation anatomique »*, alors qu'une relation *« agit sur »* sera de son côté licite entre un *« acte »* et une *« localisation anatomique »*.

La définition de ces contraintes canoniques sur les relations est un élément clé dans la mise au point d'une représentation compositionnelle des concepts d'un domaine. Elle aide à contrôler l'interprétation des relations et leur emploi pour la description de concepts composites. Le réseau sémantique du projet UMLS (Unified Medical Language System, [17]) propose ainsi une hiérarchie de types de concepts, une hiérarchie de relations, et des contraintes spécifiant quelles relations peuvent être employées avec quels concepts.

La manipulation de concepts composites, en particulier leur comparaison, nécessite des traitements plus complexes que ceux qui suffisaient avec les concepts atomiques des terminologies traditionnelles. Nous examinons ce point dans la section suivante.

4.2 Systèmes de représentation des connaissances

Nous introduisons maintenant la notion de formalisme de représentation des connaissances, dans la tradition des réseaux sémantiques (voir par exemple [18]), qui ont connu un développement important en Intelligence artificielle (IA) depuis les années 1970. Nous nous focalisons sur les représentants prototypiques de cette tradition, les logiques de description (ou logiques terminologiques [19]) et les Graphes conceptuels [20], en prenant en considération leurs principaux traits pertinents pour ce qui nous concerne ici.

4.2.1 Principes généraux

Le cœur de ces formalismes de représentation est précisément une hiérarchie de types de concepts et une hiérarchie de relations, très proches de ce que nous avons vu dans les terminologies des sections précédentes. Ces formalismes offrent un support formel à la *composition* de concepts et à leur *comparaison*. Ils ont été employés dans plusieurs entreprises de modélisation des concepts médicaux, en particulier les projets GALEN [21] et SNOMED RT [22] (voir ci-dessous à la section 4.3).

La construction de concepts composites se fait sur le mode que nous avons vu ci-dessus. Un concept complexe est formé de concepts élémentaires liés par des relations. Par exemple, notre concept *« appendicite aiguë »* peut être représenté par le

Graphe conceptuel suivant (les concepts sont encadrés de crochets, les relations de parenthèses) :

[INFLAMMATION]–
(LOCALISÉ DANS)→[APPENDICE VERMICULAIRE]
(COURS)→[AIGU] /

ou par l'expression suivante dans le langage de représentation GRAIL [23] du projet GALEN :

INFLAMMATION *qui* <
ESTLOCALISÉDANS APPENDICEVERMICULAIRE
ACOURS AIGU>

Des contraintes régissent le mode de composition des concepts et des relations pour former des concepts complexes. Ce sont les *graphes canoniques* des Graphes conceptuels ou les *restrictions de rôles* des logiques de description (ils sont encore appelés *sanctions* dans GALEN ou *modèles fondateurs* dans SNOMED RT). On pourra par exemple imposer que les concepts reliés par la relation « *localisé dans* » respectent la contrainte suivante :

[AFFECTION]→(LOCALISÉ DANS)→[LOCALISATION ANATOMIQUE]

Si le concept « *inflammation* » est bien indiqué comme étant plus spécifique que « *affection* » dans la hiérarchie de concepts, et si « *appendice vermiculaire* » est bien indiqué comme étant plus spécifique que « *localisation anatomique* », alors l'emploi de la relation « *localisé dans* » dans le concept complexe « *appendicite aiguë* » ci-dessus est considéré comme acceptable.

L'ensemble formé de la hiérarchie de concepts élémentaires, de la hiérarchie de relations et des contraintes sur leur composition, aussi appelé *support* [24], détermine l'espace des concepts qui peuvent être représentés. Il constitue de ce fait un premier niveau de *modèle conceptuel* du domaine. À la différence des terminologies précoordonnées de la section 3, l'approche compositionnelle permet de décrire une infinité de concepts à partir d'un support fini. De plus, les formalismes de représentation des connaissances discutés ici correspondent à des systèmes logiciels opératoires. Ces systèmes sont capables de vérifier automatiquement qu'un concept construit est en accord avec le support de la représentation, c'est-à-dire est constitué de concepts et relations répertoriés et combinés en respectant les contraintes énoncées. Ils savent aussi classer automatiquement des concepts composites.

4.2.2 Subsomption et classification

Un avantage décisif des formalismes de représentation des connaissances est leur capacité à comparer des concepts complexes. Déterminer si un concept est plus spécifique qu'un autre (est *subsumé* sous cet autre concept) est un élément clé pour la

classification des concepts d'un domaine. Supposons que des informations sur chaque patient ont été exprimées sur ce mode, constituant ainsi une forme de codage conceptuel des dossiers médicaux. Si l'on recherche par exemple tous les patients qui ont présenté des *« affections »* de l'*« appendice vermiculaire »*, un patient pour lequel a été entrée une *« appendicite aiguë »* devra être retrouvé. Formellement, il faut identifier tous les concepts plus spécifiques que (ou identiques à) :

[AFFECTION]→(LOCALISÉ DANS)→[APPENDICE VERMICULAIRE]

D'après les règles déterminant la subsomption, étant donné

(i) que *« inflammation »* est plus spécifique que *« affection »* dans la hiérarchie de concepts, et

(ii) que l'ajout à un concept d'une relation et d'un concept (→(COURS)→[AIGU]) spécialise ce concept,

le concept représentant notre *« appendicite aiguë »* est effectivement subsumé par le concept représentant *« affection de l'appendice vermiculaire »*.

Les systèmes de représentation des connaissances incluent des méthodes de classification automatique des concepts. La classification est effectuée systématiquement, pour chaque concept entré, dans les systèmes de la famille des logiques de description. Elle est effectuée à la demande, en employant le mécanisme de *projection*, dans ceux fondés sur les Graphes conceptuels.

L'emploi d'un formalisme de représentation des connaissances permet de décrire les informations médicales avec la granularité la plus fine compatible avec le support de la représentation : la finesse de cette granularité est bornée par celle des concepts et des relations disponibles. Grâce au mécanisme de classification, rien n'empêche en même temps d'exploiter les informations enregistrées à un grain plus grossier, par exemple pour déterminer les codes d'une classification comme la CIM. C'est ainsi que dans le projet MENELAS [25], une représentation en Graphes conceptuels des informations concernant un patient servait de base à la détermination des codes CIM (plus précisément, HCIMO) pour ce patient [26]. Les conditions de vérité de chaque code CIM pour le domaine abordé (maladies coronariennes) étaient exprimées sous forme de Graphes conceptuels, et un code CIM était assigné à un patient si le graphe représentant ce code subsumait l'un des graphes représentant les informations décrites pour ce patient.

Dans l'absolu, cette méthode peut être appliquée à partir d'une même représentation de départ pour produire des codes dans une ou plusieurs classifications différentes (par exemple, à des fins d'évaluation d'activité d'une part et d'épidémiologie d'autre part). On combine ainsi les avantages du recueil d'informations détaillées pour des besoins cliniques et du recueil d'informations agrégées pour des objectifs statistiques. Il faut noter cependant que pour que ce soit possible, il faut que le support de la représentation soit partout plus fin que la plus fine des classifications visées, et que les

informations soient enregistrées (donc saisies) à ce degré de finesse. Par ailleurs, l'emploi d'un formalisme de représentation des connaissances pour enregistrer les informations médicales ne dispense pas le concepteur du système d'information de définir les classes pertinentes pour chaque objectif spécifique de classification (par exemple, le Graphe conceptuel pour chaque code CIM dans [26]). En d'autres termes, si l'on peut se servir d'informations enregistrées dans un formalisme de représentation des connaissances pour générer automatiquement les classes d'une classification, cela ne supprime pas pour autant l'intérêt de ces classifications en tant que telles [10].

4.2.3 Ontologie

On a vu que l'épine dorsale d'un système de représentation des connaissances est son *support*. Le terme *ontologie*, issu de la philosophie de la connaissance, désigne généralement l'ensemble des concepts d'un domaine. Dans le cadre de la représentation des connaissances, ce terme est employé plus particulièrement pour décrire le contenu du support : les concepts, relations et contraintes effectivement utilisés pour modéliser un domaine donné. On peut considérer qu'une ontologie, dans ce sens, est l'aboutissement formel de la définition d'une terminologie.

La constitution d'une ontologie est un problème difficile, et de nombreux critères de bonne structuration ont été proposés (entre autres, dans le domaine médical, [27, 28, 29]). Contrairement à ce que l'on trouve dans certaines terminologies médicales, la relation hiérarchique qui structure l'ontologie doit être unique : A est fils de B signifie que A *« est-un »* B. Cela permet de définir clairement la subsomption entre concepts. À l'inverse, le MeSH ou la SNOMED utilisent alternativement les relations *« est-un »* et *« est-une-partie-de »* à différents endroits de la hiérarchie. Ces terminologies ne sont donc pas utilisables telles quelles comme ontologies. Notons qu'il est possible, sous certaines conditions, de définir une extension de la subsomption prenant en compte la relation *« est-une-partie-de »* en plus de *« est-un »* [30].

L'opposition entre *concepts primitifs* et concepts composites a des implications fondamentales dans la constitution d'une ontologie et des répercutions importantes sur les possibilités de classification multiple. Tout système formel repose sur un jeu de primitives ; un système de représentation des connaissances repose sur un jeu de concepts (et de relations) primitifs. La hiérarchie de types de concepts définit les relations de subsomption entre les concepts primitifs. Des concepts composites peuvent être définis à partir de ceux-ci, sur le mode que nous avons présenté plus haut (section 4.2.1). Comme nous l'avons vu ensuite (section 4.2.2), ces concepts composites sont automatiquement classés d'après leur définition. Selon le contenu de sa définition, un concept composite peut de plus être classé sous plusieurs autres. Ainsi, une *« appendicite aiguë »* sera classée aussi bien sous *« affection aiguë »* (comme *« infarctus aigu du myocarde »*) que sous *« affection de l'appendice vermiculaire »* (comme *« tumeur maligne de l'appendice »*).

Ce type de subsomption multiple est extrêmement pratique pour la recherche d'information comme pour le codage dans des classifications spécifiques. Dans l'exemple

examiné, la classification multiple est fondée sur des propriétés explicites des concepts composites. Pour que ce soit toujours le cas (ainsi que pour d'autres motifs formels [27]), il faut que les concepts primitifs soient hiérarchisés en arbre, c'est-à-dire que chaque concept primitif n'ait qu'un seul père [28]. Chaque concept primitif n'est donc classé qu'à un seul endroit. Cela correspond à un *point de vue* considéré comme essentiel sur ce concept. Des propriétés supplémentaires peuvent ensuite être assignées explicitement aux concepts. C'est le jeu de la subsomption sur ces propriétés explicites qui permet alors d'obtenir plusieurs classifications pour un même concept.

4.3 Des ontologies pour la médecine

Plusieurs projets ont pour but de concevoir une ontologie pour la représentation des concepts médicaux. Nous discutons brièvement les deux principaux : GALEN [21] et SNOMED RT [22]. Nous mentionnons aussi pour mémoire un projet apparenté, l'UMLS [31], et une contribution française, le projet MENELAS [25].

Le projet GALEN [21] est la première initiative d'envergure à avoir eu pour objectif la construction d'une ontologie pour la médecine. Il s'agit d'une série de deux projets européens, GALEN (1992–1994) et GALEN-IN-USE (1996–1998). La représentation employée est GRAIL [23], une variété de logique de description. Les concepts primitifs de l'ontologie de GALEN forment un arbre à quelques exceptions près. Chaque concept est accompagné d'une déclaration des relations qui doivent ou peuvent le lier à d'autres concepts. GALEN a pour but de faciliter la description d'informations cliniques, le codage et le transcodage dans des classifications diverses. Le premier projet a mené à une hiérarchie contenant de l'ordre de 4000 concepts. La version actuelle en contient bien davantage. D'après la documentation du projet, elle couvre la moitié de ce que ses auteurs estiment qu'elle doit contenir, mais avec déjà une profondeur et une complexité plus grandes que dans les terminologies existantes ; et la couverture dans certains domaines comme les actes chirurgicaux est virtuellement complète. Un formalisme intermédiaire de *dissections* [32] facilite l'entrée de concepts complexes, et un système de génération automatique d'expressions en langue naturelle [33] permet de relire des concepts sous forme d'expressions en français, anglais ou allemand.

Le projet SNOMED RT [22, 15] est un remaniement de la nomenclature SNOMED visant à épauler ses termes par des descriptions dans un langage de représentation des connaissances (K-Rep, de la famille des logiques de description). Démarré vers 1996, il est mené aux États-Unis sous forme d'une collaboration entre le College of American Pathologists (soutien institutionnel traditionnel de la nomenclature SNOMED), la société Kaiser Permanente (*Health Management Organization*, système intégré privé d'assurance maladie et de réseau de soins) et la Mayo Clinic (grand réseau de soins). L'idée est de distinguer une « terminologie de référence » (RT) des terminologies qui peuvent être utiles pour des interfaces de saisie de données ou pour des systèmes de traitement automatique des langues. En cela, ce projet est très proche des principes sous-tendant GALEN. La principale différence d'approche est que SNOMED RT part

d’une terminologie existante, la nomenclature SNOMED III. La première version de cette « terminologie de référence » est attendue pour 1999.

Le projet UMLS (*Unified Medical Language System*, [31]) n’est pas à proprement parler un projet de constitution d’ontologie. Il fait plutôt partie de la famille des terminologies, mais y tient une place particulière. Il s’agit de l’union raisonnée de plus de quarante terminologies biomédicales dont le thésaurus MeSH (y compris sa traduction en français et dans plusieurs autres langues), la Classification internationale des maladies et la nomenclature SNOMED III. Cette union est appelée *Métathésaurus*, et contenait, en 1998, 476 313 concepts et 1 051 901 termes, synonymes et autres variantes lexicales. La version 1999 contiendra plus de 625 000 concepts, et outre celle du MeSH, la traduction française de plusieurs autres terminologies internationales : *Classification internationale des soins primaires* (ICPC) et WHOART (*WHO Adverse Drug Reaction Terminology*). L’UMLS comprend aussi un *réseau sémantique* de 134 concepts (*types sémantiques*) et 54 relations (version 1999), sorte d’ontologie embryonnaire et très générale du domaine bio-médical. Dans la mesure où chaque concept du métathésaurus possède un ou plusieurs pères dans la hiérarchie de concepts du réseau sémantique, on pourrait considérer que l’ensemble formé du métathésaurus et du réseau sémantique constitue une ontologie. Dans les faits, cet ensemble ne possède pas les propriétés formelles permettant de s’en servir ainsi dans un langage de représentation des connaissances [34, 35]. L’UMLS n’en demeure pas moins une ressource précieuse pour la recherche documentaire (voir par exemple [36]).

Le projet européen MENELAS [25] (1992–1995) s’est intéressé à la construction d’une représentation formelle (dans le formalisme des Graphes conceptuels) par analyse de comptes rendus d’hospitalisation rédigés en texte libre. L’objectif était entre autres de produire automatiquement des codes CIM pour les comptes rendus analysés [26] et plus généralement de répondre à des questions concernant les informations décrites dans ces textes [37]. Au moment où MENELAS a démarré, aucune ontologie médicale n’était disponible ; le projet a donc construit sa propre ontologie, qui contenait 1800 concepts et 300 relations à la fin du projet [38]. Les principes de structuration de cette ontologie [39] se sont révélés proches de ceux employés dans le projet GALEN, mené en parallèle.

Mentionnons pour conclure cette revue que le mouvement actuel est à la convergence des différentes entreprises de modélisation des concepts médicaux (essentiellement, GALEN, SNOMED RT et les Termes Cliniques du NHS) vers l’emploi de systèmes formels proches et possiblement compatibles [40, 15].

5 Entre normalisation et expressivité

5.1 Système formel et expression naturelle

Les systèmes formels de représentation des connaissances que nous avons présentés dans la section 4.2 apportent plusieurs avantages par rapport aux terminologies

traditionnelles pour représenter l'information médicale. La compositionnalité des représentations et la possibilité de les comparer, en particulier de les classer, sont sources de puissance d'expression et de traitement. Le fait que ces systèmes soient formels rend leur comportement prédictible ; la comparaison de deux descriptions, de deux éléments d'information, de deux connaissances se fait systématiquement, de façon fiable et reproductible, en référence à l'ontologie fournie au système. Ce n'est pas toujours le cas lorsque la manipulation de termes fait appel à des connaissances externes à la structure de la terminologie elle-même, en particulier à l'interprétation de définitions en langue naturelle. La mise au gabarit d'un système formel amène à rendre explicites des connaissances qui sont considérées comme partagées dans les systèmes non formels, rendant de ce fait ces derniers dépendants de l'interprétation humaine. À l'inverse, les systèmes formels de représentation des connaissances permettent à des logiciels de manipuler des représentations de l'information médicale de façon sûre et efficiente [28].

Le prix à payer pour l'emploi d'un système formel est double. D'une part, il faut disposer d'une ontologie du domaine, dont la constitution est une tâche complexe et d'envergure pour la médecine (voir la section 4.3). D'autre part, il faut savoir saisir et pouvoir relire des descriptions formulées dans le formalisme choisi. En effet, en passant à une représentation formelle, nous nous sommes concentrés sur les *concepts*, et nous avons laissé de côté les *termes*. Il nous reste donc à examiner comment faire le lien entre l'expression *naturelle* d'une information et sa représentation formelle.

Pour ce qui concerne les concepts élémentaires (les primitives de l'ontologie), il suffira de leur faire correspondre à chacun un terme préférentiel, comme dans les terminologies de la section 3. Pour les concepts composites, la question est plus difficile. La saisie de descriptions dans un langage de représentation des connaissances n'est pas une tâche facile. Sans aide logicielle, on imagine bien la difficulté de saisir rapidement et sans erreur des formules comme celles montrées plus haut en exemple (section 4.2.1). La solution proposée par GALEN aussi bien que SNOMED RT est une saisie assistée par un *serveur terminologique* [41]. Mentionnons aussi la possibilité d'employer une représentation intermédiaire (*dissections* de GALEN [32]), plus simple, pour décrire des faits appartenant à une classe spécifique (par exemple, des actes chirurgicaux). L'idée est que pour une telle classe de faits, on peut mettre au point un modèle stéréotypé prévoyant les informations principales à fournir pour construire une représentation bien formée. Il suffit alors à l'utilisateur de spécifier ces quelques informations, qui seront ensuite converties dans le format réel de GALEN.

Une autre voie consiste à générer des représentations par analyse de textes en langue naturelle, comme les comptes rendus d'hospitalisation ou d'autres pièces du dossier patient. Le projet MENELAS (section 4.3) a montré à la fois l'intérêt et le coût de cette voie. Le succès du système MedLEE d'analyse de comptes rendus de radiologie [42], qui fonctionne en routine depuis trois ans, semble cependant indiquer que ce type d'approche est déjà pratiquable dans des conditions spécifiques : si on l'applique à des textes « plus simples », comme des comptes rendus de radiologie, en visant des représentations moins élaborées qu'un formalisme de représentation des connaissances.

Notons qu'il est également question d'étendre le système d'analyse RECIT [43] pour produire des représentations basées sur l'ontologie de GALEN.

À l'inverse, une description étant disponible dans un formalisme de représentation des connaissances, son examen par une personne n'est pas toujours très informatif : ces représentations deviennent rapidement complexes et absconses. La méthode la plus universellement utile consiste à générer à partir d'un concept, simple ou complexe, une expression en langue naturelle. C'est ce que permet le générateur associé à GALEN [33]. Notons que ce générateur a aussi été employé pour valider des représentations mises au point pour modéliser la nouvelle nomenclature des actes médicaux français [44]. Une autre possibilité consiste à produire à partir d'une représentation donnée les codes d'une *terminologie traditionnelle* comme la CIM ou la SNOMED.

5.2 Mot, terme, concept : approche normative ou approche descriptive ?

Les problèmes de correspondance entre termes et concepts réveillent des interrogations sur les hypothèses qui sous-tendent l'entreprise terminologique présentée à la section 3, dont le passage à un langage de représentation des connaissances (section 4) est l'aboutissement formel.

Il existe une tension entre deux forces antagonistes en traitement de l'information médicale. La première est la nécessité de normalisation, qui conditionne l'échange d'information. Elle correspond à une *approche normative* de la représentation de l'information médicale. La seconde force est mue par le besoin d'expressivité, d'adaptation à l'évolution constante des connaissances et des techniques en médecine. Elle demande une *approche descriptive* des notions effectivement maniées dans la pratique médicale, en particulier pour le soin aux patients, et une meilleure prise en compte de la façon dont ces notions sont exprimées dans les documents textuels.

Nous soulignions à la section 2 les problèmes issus de la non-compositionnalité de la langue. Ces problèmes sont liés à la *contextualité* du sens : le sens des mots varie selon leur contexte d'emploi. Un mot, une expression devient un terme lorsque, par convention, on lui attribue une signification indépendante des variations contextuelles (et temporelles) : lorsqu'on le *décontextualise* [9]. Le terme est ainsi *« un artefact de la discipline qui l'instaure »* (ibid). De plus, le fait que des terminologues proposent de considérer, dans une discipline donnée, une expression comme un terme, fait émerger et normalise un concept associé : *« un concept n'est pas la source du terme, mais le produit de son instauration »* (ibid). La normalisation des concepts est donc le résultat d'un travail sur la langue.

On peut en conclure que la mise au point de terminologies et d'ontologies doit s'appuyer sur la linguistique, en étudiant les textes spécialisés [45, 46, 47]. Comptes rendus médicaux, manuels de cours, articles scientifiques témoignent des notions effectivement maniées par les acteurs de la médecine. C'est en les observant que l'on peut concevoir ou tenir à jour terminologies et ontologies « normalisées ». Les travaux ré-

cents témoignent d'un intérêt croissant pour ce type d'approche [48, 49, 50, 51, 52, 53].

Références

[1] Musen MA et van Bemmel JH. *Handbook of Medical Informatics*. Springer-Verlag, 1997.

[2] Degoulet P et Fieschi M. *Informatique médicale*. Abrégés. Masson, Paris, 1994.

[3] Huff S. Clinical data exchange standards and vocabularies for messages. *J Am Med Inform Assoc* 1998;5(suppl).

[4] Degoulet P, Sauquet D, Jaulent MC, Zapletal E, et Lavril M. Rationale and design considerations for a semantic mediator in health information systems. *Methods Inf Med* 1998;37(4–5):518–26.

[5] Nazarenko A. Le principe de compositionnalité sémantique : un enjeu pour le traitement automatique des langues. *Traitement Automatique des Langues* 1998;39(1):3–7. Présentation du numéro spécial *Compositionnalité*.

[6] Lerat P. *Les langues spécialisées*. Presses Universitaires de France, 1995.

[7] Otman G. Pourquoi parler de connaissances terminologiques et de bases de connaissances terminologiques. *La banque des mots* 1994;6:5–28.

[8] Scherrer JR. Concepts, knowledge and language information systems: Follow-up 30 months later. *Methods Inf Med* 1998;37(4–5):312–4.

[9] Rastier F. Le terme : entre ontologie et linguistique. *La banque des mots* 1995;7:35–65.

[10] Ingenerf J et Giere W. Concept-oriented standardization and statistics-oriented classifications: Continuing the classification versus nomenclature controversy. *Methods Inf Med* 1998;37(4–5):527–39.

[11] Rossi Mori A, Consorti F, et Galeazzi E. Standards to support development of terminological systems for healthcare telematics. *Methods Inf Med* 1998;37(4–5):551–63.

[12] Medical Subject Headings. WWW page `http://www.nlm.nih.gov/mesh/meshhome.html`, National Library of Medicine, Bethesda, Maryland, 1998.

[13] Organisation mondiale de la Santé, Genève. Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision, 1993.

[14] Côté RA, Rothwell DJ, Palotay JL, Beckett RS, et Brochu L, eds. *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. College of American Pathologists, Northfield, 1993.

[15] Spackman K et Campbell K. Compositional concept representation using SNOMED: Towards further convergence of clinical terminologies. *J Am Med Inform Assoc* 1998;5(suppl).

[16] Brown PJB, O'Neil M, et Price C. Semantic definition of disorders in version 3 of the Read codes. *Methods Inf Med* 1998;37(4–5):415–9.

[17] McCray AT. The UMLS semantic network. In: Proc Thirteenth Annu Symp Comput Appl Med Care, Washington. IEEE, 1989:503–7.

[18] Sowa JF, ed. *Principles of Semantic Networks*. Morgan Kaufmann Publishers, San Mateo, Ca., 1991.

[19] Brachman RJ et Schmolze J. An overview of the KL-ONE knowledge representation system. *Cogn Sci* 1985;9:171–216.

[20] Sowa JF. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, London, 1984.

[21] Rector AL, Nowlan WA, et Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Piemme T, et Rienhoff O, eds, Proc MEDINFO 92, Geneva. North Holland, 1992:1420–6.

[22] Spackman K, Campbell K, et Côté RA. SNOMED RT: A reference terminology for health care. *J Am Med Inform Assoc* 1997;4(suppl):640–4.

[23] Rector AL, Bechhover S, Goble CA, et al. The GRAIL concept modelling language for medical terminology. *Artif Intell Med* 1997;9(2):139–71.

[24] Chein M et Mugnier ML. Conceptual Graphs: fundamental notions. *Rev d'Intell Artif* 1992;6(4):365–406.

[25] Zweigenbaum P et Consortium MENELAS. MENELAS: an access system for medical records using natural language. *Comput Methods Programs Biomed* 1994;45:117–20.

[26] Delamarre D, Burgun A, Seka LP, et Le Beux P. Automated coding system of patient discharge summaries using Conceptual Graphs. *Methods Inf Med* 1995;34:345–51.

[27] Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, et Boisvieux JF. Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods Inf Med* 1995;34(1/2):15–24.

[28] Rector AL. Thesauri and formal classifications: Terminologies for people and machines. *Methods Inf Med* 1998;37(4–5):501–9.

[29] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;37(4–5):394–403.

[30] Bernauer J et Goldberg H. Compositional classification based on Conceptual Graphs. In: Andreassen et al. S, ed, Proc Artificial Intelligence in Medicine Europe, Munich. IOS Press, 1993:348–59.

[31] Lindberg DAB, Humphreys BL, et McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32(2):81–91.

[32] Rogers J, Solomon WD, Rector AL, Zanstra P, et van der Haring EJ. From rubrics to dissections to GRAIL to classifications. In: Pappas C, Maglaveras N, et Scherrer JR, eds, Proceedings of MIE'97, Thessaloniki, Grece. IOS Press, 1997.

[33] Wagner J, Solomon W, Michel P, et al. Multilingual natural language generation as part of a medical terminology server. In: Greenes RA, Peterson HE, et Protti DJ, eds, Proc 8 th World Congress on Medical Informatics, 1995:100–4.

[34] Carenini G et Moore JD. Using the UMLS semantic network as a basis for constructing a terminological knowledge base: A preliminary report. In: Proc Seventeenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1993:725–9.

[35] Volot F, Zweigenbaum P, Bachimont B, et al. Structuration and acquisition of medical knowledge: Using UMLS in the Conceptual Graph formalism. In: Proc Seventeenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1993:710–4.

[36] Joubert M, Fieschi D, Fieschi M, et Volot F. Conceptual integration of information databases into an intranet. In: Cesnik B, Safran C, et Degoulet P, eds, Proc 9 th World Congress on Medical Informatics, 1998.

[37] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, et Boisvieux JF. Évaluation d'une représentation conceptuelle normalisée de comptes rendus médicaux en langue naturelle. In: Proceedings of the 11 th Conference RFIA-AFCET, Clermont-Ferrand, France. AFCET, janvier 1998:III.261–270.

[38] Zweigenbaum P et Consortium MENELAS. MENELAS: coding and information retrieval from natural language patient discharge summaries. In: Laires MF, Ladeira MJ, et Christensen JP, eds, *Advances in Health Telematics*. IOS Press, Amsterdam, 1995:82–9. MENELAS Final Edited Progress Report.

[39] Bouaud J, Bachimont B, Charlet J, et Zweigenbaum P. Methodological principles for structuring an "ontology". In: IJCAI'95 Workshop on "Basic Ontological Issues in Knowledge Sharing", août 1995.

[40] Chute CG. The Copernican era of healthcare terminology: A re-centering of health information systems. *J Am Med Inform Assoc* 1998;5(suppl).

[41] Rector AL, Solomon WD, Nowlan WA, et Rush TW. A terminology server for medical language and medical information systems. *Methods Inf Med* 1995;34(1/2).

[42] Friedman C, Alderson PO, Austin JH, Cimino JJ, et Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1(2):161–74.

[43] Rassinoux AM, Wagner JC, Lovis C, et al. Analysis of medical texts based on a sound medical model. *J Am Med Inform Assoc* 1995;2(suppl):27–31.

[44] Rodrigues JM, Trombert-Paviot B, Baud R, Wagner J, et Meusnier F. Galen-In-Use: Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. In: Cesnik B, Safran C, et Degoulet P, eds, Proc 9th World Congress on Medical Informatics, 1998.

[45] Condamines A. Terminologie et représentation des connaissances. *La banque des mots* 1994;6:29–44.

[46] Bourigault D et Condamines A. Réflexions sur le concept de base de connaissances terminologiques. In: Actes des Cinquièmes journées nationales du PRC-GDR IA, Nancy. 1995.

[47] Biébow B et Szulman S. Méthodologie de création d'un noyau de base de connaissances en logique terminologique à partir de textes. In: Actes 2e rencontres Terminologie et intelligence artificielle, Toulouse. ERSS, avril 1997.

[48] Bourigault D. Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. In: RFIA'94. AFCET, 1994:1123–32.

[49] Jacquemin C. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches, Université de Nantes, 1997.

[50] Hersh WR, Campbell EH, Evans DA, et Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. *J Am Med Inform Assoc* 1996;3(suppl):159–63.

[51] Chute CG et Elkin PL. A clinically derived terminology: Qualification to reduction. *J Am Med Inform Assoc* 1997;4(suppl).

[52] Nazarenko A, Zweigenbaum P, Bouaud J, et Habert B. Corpus-based identification and refinement of semantic classes. *J Am Med Inform Assoc* 1997;4(suppl):585–9.

[53] Nelson SJ, Kuhn T, Radzinski D, et al. Creating a thesaurus from text: A "bottom-up" approach to organizing medical knowledge. *J Am Med Inform Assoc* 1998;5(suppl).